A Generalized Cloud Storage Architecture with Backup Technology for any Cloud Storage Providers

Talasila Sasidhar[1], Pavan Kumar Illa[2], Subrahmanyam Kodukula[3]

[1]PG Scholar, K L University, Andhra Pradesh, India, sasidhar369@gmail.com
[2]PG Scholar, K L University, Andhra Pradesh, India, illa.pavankumar@gmail.com
[3]Professor, K L University, Andhra Pradesh, India, smkodukula@yahoo.com

ABSTRACT

Cloud Storage Architecture is major topic in now a day because the data usage and the storage capacity are increased double year by year. So that some of the major companies are mainly concentrated on demand storage option like cloud storage. The existing cloud storage providers are mainly concentrated on performance, cost issues and multiple storage options. In this paper we discussed about the use of the Backup Technology when it is integrated in the cloud architecture. In this Backup Technology two major backup options Snapshotting and Disaster Recovery are to be discussed. The major intention is providing a new architecture that is useful for further research in Cloud Storage Architectures.

Keywords: Cloud Storage Architecture, Snapshotting, Disaster Recovery

INTRODUCTION

Cloud storage architectures are primarily about delivery of storage on demand in a highly scalable and multi-tenant way. Generically (see Figure 1), cloud storage architectures consist of a front end that exports an API to access the storage. In traditional storage systems, this API is the SCSI protocol; but in the cloud, these protocols are evolving. There, you can find Web service front ends, file-based front ends, and even more traditional front ends (such as Internet SCSI, or iSCSI).Behind the front end is a layer of middleware that I call the storage logic. This layer implements a variety of features, such as replication and data reduction, over the traditional data-placement algorithms (with consideration for geographic placement).Finally, the back end implements the physical storage for data. This may be an internal protocol that implements specific features or a traditional back end to the physical disks [1].

Cloud Storage Characteristics

Manageability

One key focus of cloud storage is cost. If a client can buy and manage storage locally compared to leasing it in the cloud, the cloud storage market disappears. But cost can be divided into two high-level categories: the cost of the physical storage ecosystem itself and the cost of managing it. The management cost is hidden but represents a long-term component of the overall cost. For this reason, cloud storage must be self-managing to a large extent. The ability to introduce new storage where the system automatically self-configures to accommodate it and the ability to find and self-heal in the presence of errors is critical. Concepts such as autonomic computing will have a key role in cloud storage architectures in the future [1].

Access method

One of the most striking differences between cloud storage and traditional storage is the means by which it's accessed. Most providers implement multiple access methods, but Web service APIs are common. Many of the APIs are implemented based on REST principles,

which imply an object-based scheme developed on top of HTTP (using HTTP as a transport). REST APIs are stateless and therefore simple and efficient to provide. Many cloud storage providers implement REST APIs, including Amazon Simple Storage Service (Amazon S3), WindowsAzure™, and Mezeo Cloud Storage Platform.

One problem with Web service APIs is that they require integration with an application to take advantage of the cloud storage. Therefore, common access methods are also used with cloud storage to provide immediate integration. For example, file-based protocols such as NFS/Common Internet File System (CIFS) or FTP are used, as are block-based protocols such as iSCSI. Cloud storage providers such as Nirvanix, Zetta, and Cleversafe provide these access methods [1].

Performance

There are many aspects to performance, but the ability to move data between a user and a remote cloud storage provider represents the largest challenge to cloud storage. The problem, which is also the workhorse of the Internet, is TCP.

TCP controls the flow of data based on packet acknowledgements from the peer endpoint. Packet loss, or late arrival, enables congestion control, which further limits performance to avoid more global networking issues. TCP is ideal for moving small amounts of data through the global Internet but is less suitable for larger data movement, with increasing round-trip time (RTT).

Amazon, through Aspera Software, solves this problem by removing TCP from the equation. A new protocol called the Fast and Secure Protocol (FASP™) was developed to accelerate bulk data movement in the face of large RTT and severe packet loss. The key is the use of the UDP, which is the parter transport protocol to TCP. UDP permits the host to manage congestion, pushing this aspect into the application layer protocol of FASP

Multi-tenancy

One key characteristic of cloud storage architectures is called multi-tenancy. This simply means that the storage is used by many users (or multiple "tenants"). Multi-tenancy applies to many layers of the cloud storage stack, from the application layer, where the storage namespace is segregated among users, to the storage layer, where physical storage can be segregated for particular users or classes of users. Multi-tenancy even applies to the networking infrastructure that connects users to storage to permit quality of service and carving bandwidth to a particular user [2].

Scalability

You can look at scalability in a number of ways, but it is the on-demand view of cloud storage that makes it most appealing. The ability to scale storage needs (both up and down) means improved cost for the user and increased complexity for the cloud storage provider. Scalability must be provided not only for the storage itself (functionality scaling) but also the bandwidth to the storage (load scaling).

Another key feature of cloud storage is geographic distribution of data (geographic scalability), allowing the data to be nearest the users over a set of cloud storage data centers (via migration). For read only data, replication and distributions are also possible.

Once a cloud storage provider has a user's data, it must be able to provide that data back to the user upon request. Given network outages, user errors, and other circumstances, this can be difficult to provide in a reliable and deterministic way.

Control

A customer's ability to control and manage how his or her data is stored and the costs associated with it is important. Numerous cloud storage providers implement controls that give users greater control over their costs.

Amazon implements Reduced Redundancy Storage (RRS) to provide users with a means of minimizing overall storage costs. Data is replicated within the Amazon S3infrastructure, but with RRS, the data is replicated fewer times with the possibility for data loss. This is ideal for data that can be recreated or that has copies that exist elsewhere. Nirvanix also provides policy-based replication to enable more granular control over how and where data is stored.

Efficiency

Storage efficiency is an important characteristic of cloud storage infrastructures, particularly with their focus on overall cost. The next section speaks to cost specifically, but this characteristic speaks more to the efficient use of the available resources over their cost.

To make a storage system more efficient, more data must be stored. A common solution is data reduction, whereby the source data is reduced to require less physical space. Two means to achieve this include compression—the reduction of data through encoding the data using a different representation—and de-duplication—the removal of any identical copies of data that may exist. Although both methods are useful, compression involves processing (re-encoding the data into and out of the infrastructure), where de-duplication involves calculating signatures of data to search for duplicates.

Cost

One of the most notable characteristics of cloud storage is the ability to reduce cost through its use. This includes the cost of purchasing storage, the cost of powering it, the cost of repairing it (when drives fail), as well as the cost of managing the storage. When viewing cloud storage from this perspective (including SLAs and increasing storage efficiency), cloud storage can be beneficial in certain use models [2].

An interesting peak inside a cloud storage solution is provided by a company called Backblaze (see Resources for details). Backblaze set out to build inexpensive storage for a cloud storage offering. A Backblaze POD (shelf of storage) packs 67TBin a 4U enclosure for under US$8,000. This package consists of a 4U enclosure, a motherboard, 4GB of DRAM, four SATA controllers, 45 1.5TB SATA hard disks, and two power supplies.

On the motherboard, Backblaze runs Linux® (with JFS asthe file system) and GbE NICs as the front end using HTTPS and Apache Tomcat. Backblaze's software includes de-duplication, encryption, and RAID6 for data protection.

Backblaze's description of their POD (which shows you in detail how to build your own) shows you the extent to which companies can cut the cost of storage, making cloud storage a viable and cost-efficient option.

TRADITIONAL STORAGE VS CLOUD STORAGE

In traditional storage the multiple Storage Options, for memory we have caches, RAM disks, for DAS we use local block devices (disks), for SAN / NAS types we use network-attached block devices (LUNs) / file systems (NFS & CIFS fileservers), and for Message Queues we widely use FIFOs. But where as in cloud storage if we taking the AWS services as example, for memory they use Amazon Elastic Cache, for structural storage they sue EC2 Database AMIs or Amazon SimpleDB, whereas for message queues the provide Amazon Simple Queue Service(SQS), and finally for backup the use EBS Snapshots.

Cloud Storage Options Provided By Amazon Web Services

AWS offers multiple cloud-based storage options. Each has a unique combination of performance, durability, cost, and interface, and is further enhanced by additional factors such as elasticity, availability, and scalability. These additional factors are critical for web-scale cloud-based solutions. As with traditional on-premise applications, you can use multiple cloud storage options together to form a comprehensive data storage hierarchy [3].

The primary data storage options available with the AWS cloud computing platform are [7].

- Amazon EC2 Elastic Block Storage (EBS) volumes
- Amazon EC2 Local Instance Store (Ephemeral) volumes
- Amazon Simple Storage Service (Amazon S3)
- Amazon Simple Queue Service (SQS)
- Amazon SimpleDB
- Amazon EC2 Relational Databases
- Amazon Relational Database Service (RDS)

Amazon Elastic Block Store (EBS) Volumes

Amazon Elastic Block Store (EBS) Volumes provide durable block-level storage for use with Amazon EC2 instances (virtual machines). Amazon EBS volumes are off-instance, network-attached storage that persists independently from the running life of a single Amazon EC2 instance. After an EBS volume has been attached to an Amazon EC2 instance, you are free to interact with it just as you would a physical hard disk drive, typically by formatting it with a file system of your choice. You can use an EBS volume to boot an Amazon EC2 instance (EBS AMIs only), and attach multiple EBS volumes to a single Amazon EC2 instance. Note, however, that any single EBS volume may be attached to only one Amazon EC2 instance at any point in time. An EBS volume cannot be shared with other users, unless you create an EBS snapshot (see the following "Durability and Availability" section). Sizes for EBS volumes range from 1 GB to 1 TB, and are allocated in 1GB increments.

A. Ideal Usage Scenario

Amazon EBS is meant for data that changes relatively frequently and requires long-term persistence. EBS provides persistent virtual block mode storage for Amazon EC2 virtual servers, so you can use it just as you would use a hard drive on a physical server. Amazon EBS is particularly well-suited for use as the primary storage for a file system, database or for any applications that require fine granular updates and access to raw, unformatted block-level storage [3].

B. Performance

In general, you can expect individual EBS volumes to have performance, mean time to failure (MTTF), and reliability comparable to an externally powered USB drive. Note that while EBS volumes appear as local disk drives, they are actually network-attached to an Amazon EC2 instance. Therefore, other network I/O performed by the instance, as well as the total load on the shared network, can affect individual EBS volume performance.

While each application (and its associated performance) is unique, you are free to design and deploy many traditional disk throughput optimization techniques with EBS volumes. The combination of Amazon EC2 and EBS enables you to use many of the same performance optimization techniques that you use with on-premise servers and storage. For example, you could create several volumes and then attach them all to a single Amazon EC2 instance.

With multiple EBS volumes attached you can then partition the total application I/O load by allocating one volume for log data, one volume for the database, and yet another volume for file data. Alternatively, you could stripe your data across multiple EBS volumes using a software RAID 0 device driver, thus aggregating available IOPs, total volume throughput, and total volume size.

C. Durability and Availability

Each Amazon EBS volume is automatically replicated within the same Availability Zone to prevent data loss due to failure of any single hardware component. Amazon EBS also provides the ability to create point-in-time snapshots of volumes, which are persisted to Amazon S3 (see below). These snapshots can be used as the starting point for new Amazon EBS volumes, and protect data for long-term durability.

The durability of your EBS volume depends on both the size of your volume and the amount of data that has changed since your last snapshot. EBS snapshots are incremental, point-in-time backups, containing only the data blocks changed since the last snapshot. EBS volumes that operate with 20GB or less of modified data since their most recent snapshot can expect an annual failure rate (AFR) between 0.1% - 0.5%. . In order to maximize both durability and availability of data stored in EBS volumes, users should snapshot their EBS volumes frequently. In the event that your Amazon EBS volume does fail, all snapshots of that volume will remain intact, and will allow you to recreate your volume from the last snapshot point [3].

Amazon EBS volumes are designed to be highly available. However, because EBS volumes are created in a particular Availability Zone, they will be unavailable if the Availability Zone itself is unavailable. Note that while any single EBS volume is constrained to single Availability Zone, an EBS snapshot of a volume is available across all the Availability Zones within a Region, and you can use an EBS snapshot to create one or more new EBS volumes in any Availability Zone. EBS snapshots can also be shared with other user accounts. This provides an easy-to-use “disk clone” and “disk image” backup and sharing mechanism. In order to maximize both durability and availability of their EBS data, users should snapshot their EBS volumes frequently.

D. Cost

As with all Amazon Web Services, with Amazon Elastic Block Store you pay only for what you use, with no minimum fees or long-term contracts. Amazon EBS is priced per GB-month of provisioned storage and per million I/O requests. Volume storage is charged by the amount you allocate until you release it. Amazon EBS Snapshots are priced per GB-month of data stored, as well as per 1,000 PUT requests and per 10,000 GET requests when saving and loading snapshots. For EBS snapshots, you are charged only for storage actually used (consumed). Note that EBS snapshots are incremental and compressed, so the storage used in any snapshot is generally much less than the storage consumed on an EBS volume [3].

CLOUD STORAGE USE CASE

Some enterprise environments manage two disparate sets of information. Table-oriented data is maintained in an on-premise Oracle database, while a SAN is used as a repository for file-based information. For further safeguarding of these vital assets, tapes are used for backup and disaster recovery purposes. Approximately 20 GB of new information is generated each day. Unfortunately, the backup and archive management processes are cumbersome and expensive, while restoring archived information can take days to complete.

In evaluating all potential solutions, the IT team is motivated to boost the reliability of its new storage architecture over the archaic tape system now being used. The speed of archiving and restoring data is a big factor, while cost is also a major determinant. Finally, any new arrangement must be able to work with their already-deployed storage management software. In this situation, the optimal cloud-based AWS storage architecture would employ Amazon S3 as the destination for both file-based and relational data [4].

PROBLEMS IN EXISTING SYSTEM

S3 bucket-style file storage

Amazon created this category with their S3 service that allows you to add and read files in a type of file system that they call "buckets". Other vendors have followed. You can use various API's and protocol to put and get files, including HTTP get direct from storage. It is big, expandable storage, and it is highly available on the Internet, and it is cheap. This type of storage has a simplification that makes it almost, but not quite like, a file system. You can add files, replace files, and read files, but you can't modify the files. The vendor can use home-grown caching, layering, and redundancy without having to worry about locking any single version of the file. It's great for photos, videos, messages, message attachments, document repositories, and backup. On a byte count measure, it probably will dominate Internet storage. It's not useful for databases, repositories, indexes, or other systems that update, append, and modify files [5].

Single-mount storage

Amazon offers "Elastic Block Store", and many of the new cloud vendors offer even more integrated storage for your virtual servers. This is mounted like a local disk, but it is stored on a SAN or fileserver somewhere. If you need to restart your virtual server, it gets reattached automatically (in the integrated version) or manually (in the EBS version). This is a nice hosted version of a traditional hard disk, and it will satisfy most storage needs. It is what the cloud market is providing now.

These network-mounted volumes have the advantage of using RAID and/or SAN for underlying storage, so they presumably have redundancy and seldom need any backup or restore operations [5]. But, in this case we need to the backup and restore plan if the underlying storage device fails. Don't just take it on faith that it will be well managed or rapidly restored, because some of these systems use file servers that can fail. We may find that we need an external backup, and this will introduce the long restore times. So the biggest problem about this type of storage is size limits. For example, on Amazon you can get a volume up to 1 TB in size, and it can be mounted on one virtual computer. If you have more than 1 TB, you are going to do a lot of work to allocate files between multiple servers.

However, storage failures are more difficult to manage. You might try failing over to a different server, but if that server is using the same network attached storage system, it will also be slow or stopped. You can replicate to other storage systems, but that gives you a tradeoff. The replication uses network IO, so you get the storage bandwidth problems even more frequently [6].

NEW GENERALIZED CLOUD STORAGE ARCHITECTURE

Cloud storage providers are providing various storage options and they mainly focus on storage logic and this storage is on demand as well as elastic so the cost is calculated as per usage. But if we consider the backup technology they treated as separate service. If we want to back up huge amount of data some technical synchronization issues are arise.

So that I include the backup technology in general cloud storage architecture only as shown in Figure 1. The advantage of this type of architecture is we can easily monitor the Snapshotting mechanisms and Disaster Recovery, so the synchronization problems are solved when we taking backup rapidly.

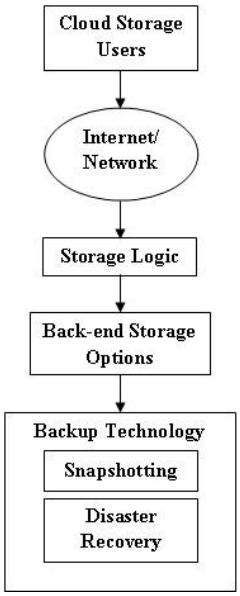

Fig 1: A Generalized Cloud Storage Architecture

CONCLUSION

In this paper we mainly focus on backup technology that is used by the existing cloud providers and the disadvantages of this system, normally all these cloud storage providers are considered this one as separate service and not included in the normal cloud storage procedure. So that for the better results and better security for the user's databases or normal data we need backup servers and backup technology that is included in the general structure only for avoiding the synchronization problems. In general these snapshotting and disaster recovery methods are followed by all the cloud storage providers, but in this proposed system we will take all these methods under one roof. For the future development of any type of cloud storage architectures this will be helpful.

REFERENCES

[1] M. Tim Jones, "Anatomy of a cloud storage infrastructure-Models, features, and internals", *IBM DeveloperWorks*, 2010.

[2] http://www.ibm.com/developerworks/cloud/ library/cl-cloudstorage/

[3] Joseph Baron, Robert Schneider, "Storage Options in the AWS Cloud", *Amazon White Papers*, Dec 2010.

[4] Joseph Baron, Robert Schneider, "Storage Options in the AWS Cloud: Use Cases", *Amazon White Papers*, Dec 2010.

[5] http://blog.assembla.com/assemblablog/tabid/12618/bid/12155/Terabytes-on-demand-Cloud-storage-options.aspx

[6] http://blog.assembla.com/assemblablog/tabid/12618/bid/44389/Problems-with-Amazon-EC2-is-storage-architecture.aspx

[7] Amazon Web Services: aws.amazon.com/